(12) UK Patent Application (19) GB (11) 2 358 760 (13) A

(43) Date of A Publication 01.08.2001

(21) Application No 0001654.3

(22) Date of Filing 25.01.2000

(71) Applicant(s)
3Com Corporation
(Incorporated in USA - Delaware)
5400 Bayfront Plaza, Santa Clara,
California 95952-8145, United States of America

(72) Inventor(s)
Michael Joseph Gleeson
John Hickey
Paul Linehan
Kevin Loughran

(74) Agent and/or Address for Service
Bowles Horton
Felden House, Dower Mews, High Street,
BERKHAMSTED, Herts, HP4 2BL, United Kingdom

(51) INT CL[7]
H04L 12/56 // H04Q 11/04

(52) UK CL (Edition S)
H4K KTKX

(56) Documents Cited
GB 2337674 A GB 2337659 A EP 0835009 A

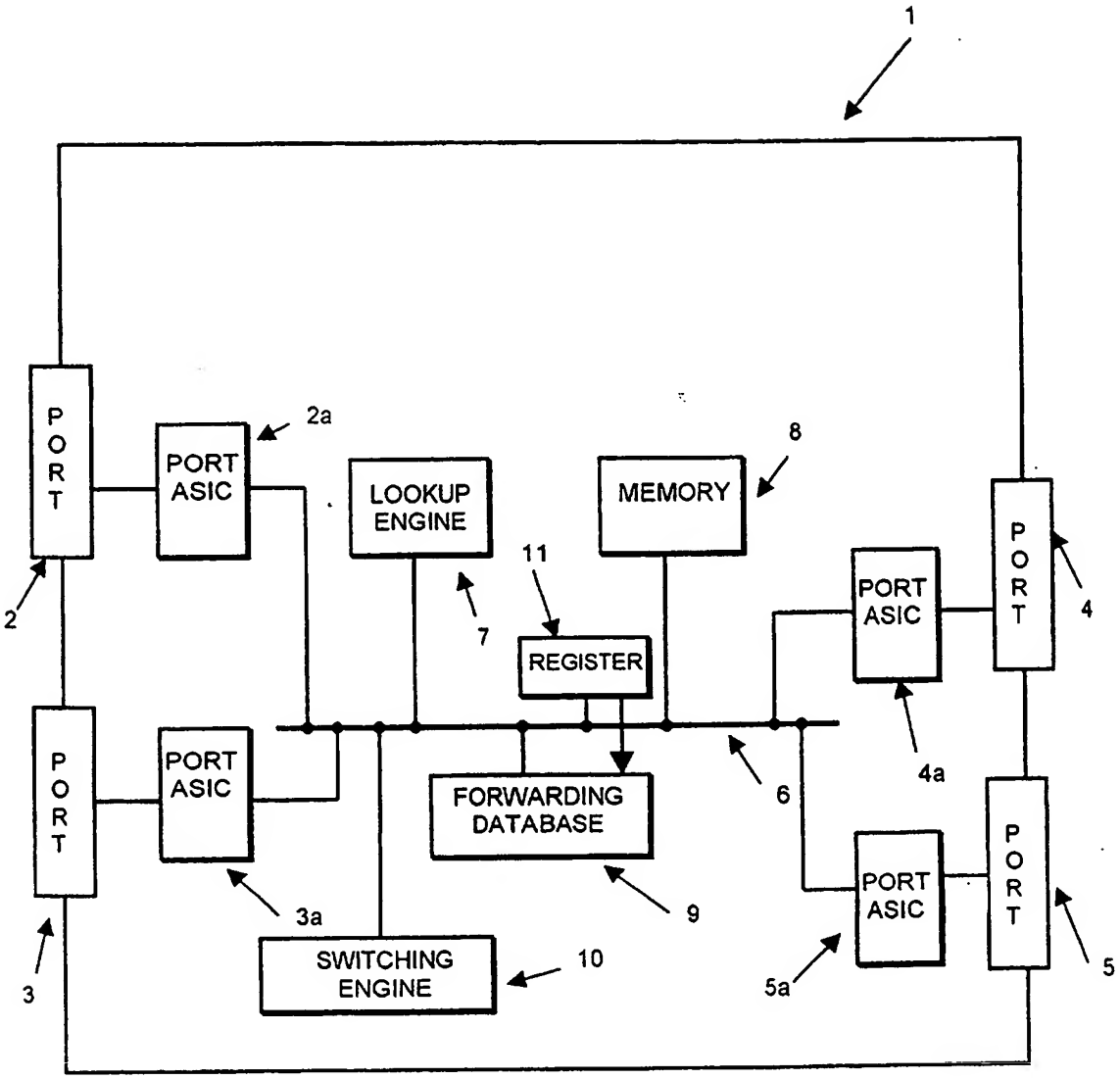
(58) Field of Search
UK CL (Edition R) H4K KTK , H4P PPS
INT CL[7] H04L 12/56 , H04Q 11/04
ONLINE : WPI ; EPODOC ; JAPIO

(54) Abstract Title
Network switch with self-learning routing facility

(57) A sub-system for forming part of a packet-data communication system for the transmission of data packets which contain media access control addresses and protocol addresses. The sub-system comprises a multiple-port network switch for forwarding data packets in response to an examination of the packets on the basis of the packets protocol addresses, and for performing said examination when a destination address in the packet corresponds to the media access control address of the switch, the switch being arranged for forwarding a packet on the basis of a media access control address therein when the destination address of the packet does not correspond to the media access control address of the switch, said switch including a forwarding database which comprises entries which relate a protocol address with a media access control address and an identification of a port of the switch. This switch may perform routing within a local-area network. The switch is configured so that protocol addresses may not be learned in the database in respect of said particular port. A network router, which may perform routing functions in respect of a wide-area network, is connected to the switch only by way of said particular port. The sub-system enables the switch to off-load from the router routing functions for the local-area network.

GB 2 358 760 A

FIG.1


1
PORT
PORT ASIC
2a
LOOKUP ENGINE
MEMORY
8
11
REGISTER
7
2
PORT ASIC
PORT
4
4a
PORT
PORT ASIC
3a
FORWARDING DATABASE
6
9
PORT ASIC
PORT
3
SWITCHING ENGINE
10
5a
5

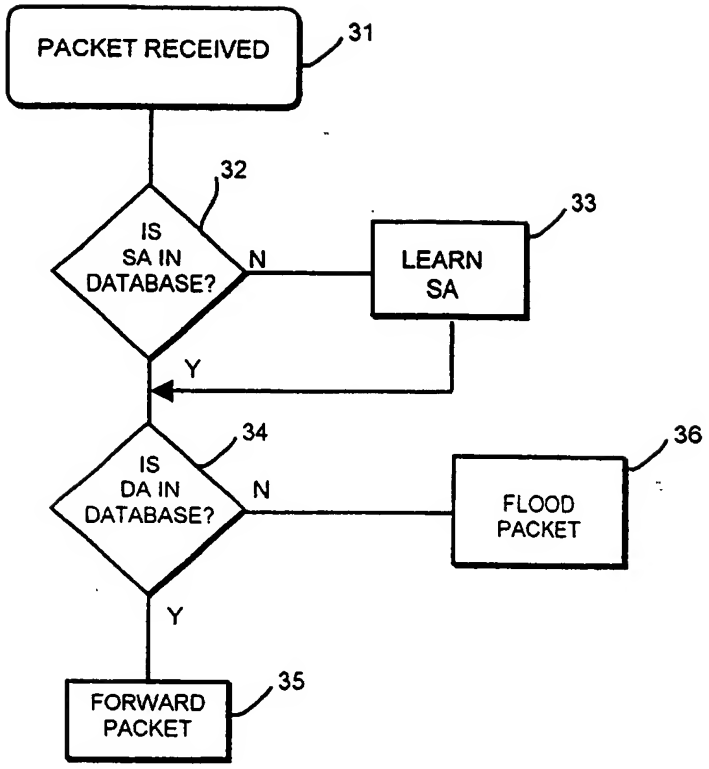

PACKET RECEIVED
31
32
IS SA IN DATABASE?
N
33
LEARN SA
Y
34
IS DA IN DATABASE?
N
36
FLOOD PACKET
Y
FORWARD PACKET
35

FIG.3

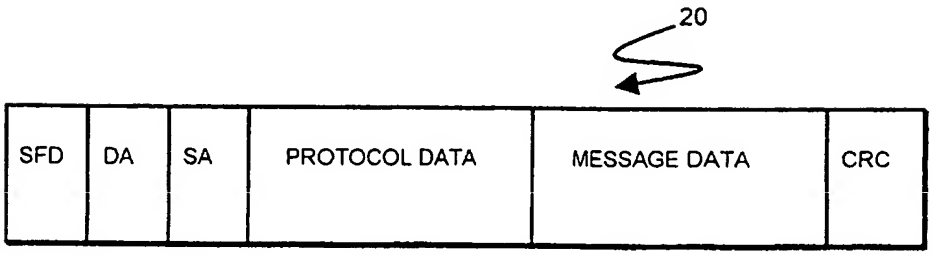

20
SFD
DA
SA
PROTOCOL DATA
MESSAGE DATA
CRC

FIG.2

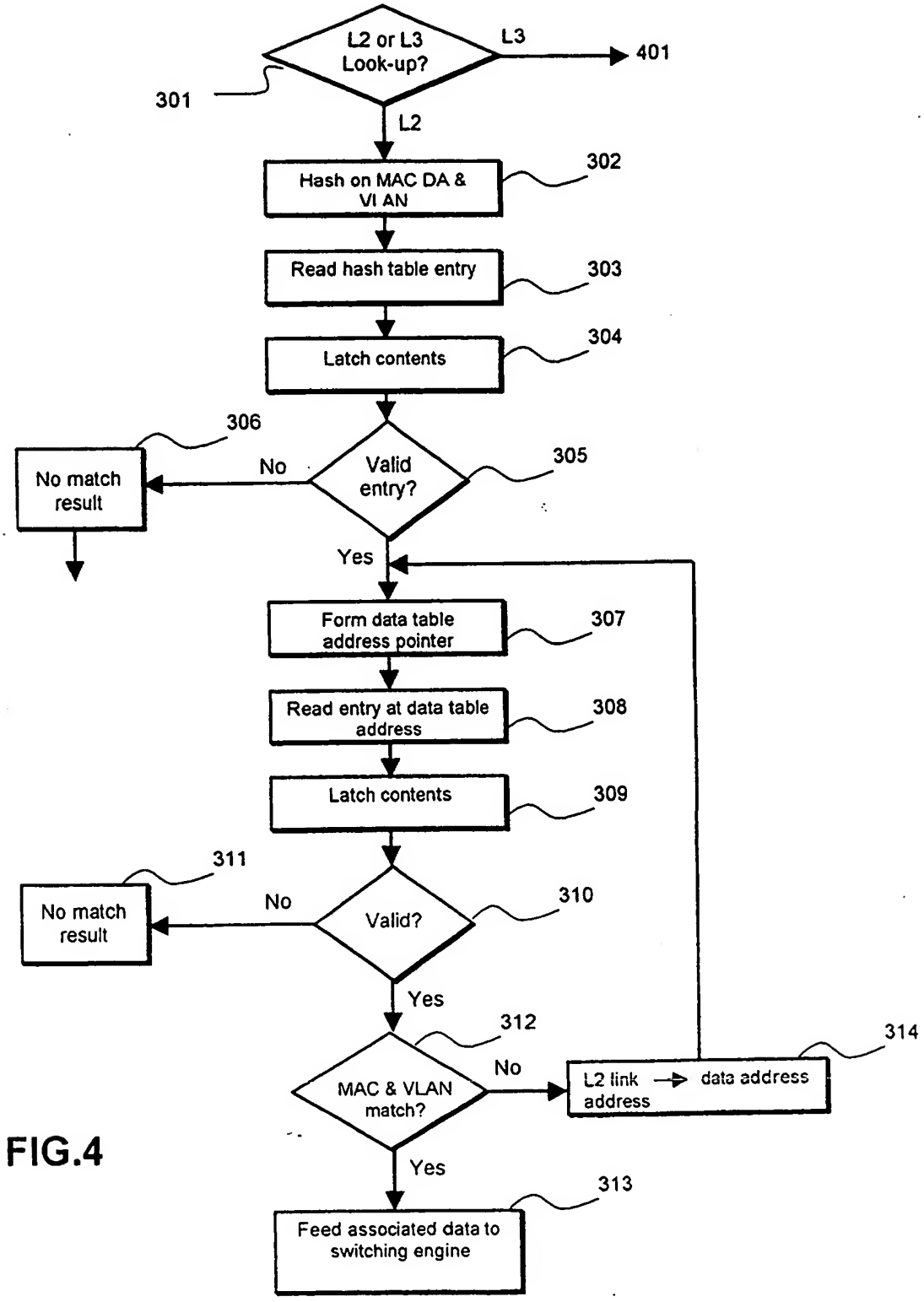

L2 or L3 Look-up?
L3
401
301
L2
Hash on MAC DA & VI AN
302
Read hash table entry
303
Latch contents
304
306
No match result
No
Valid entry?
305
Yes
Form data table address pointer
307
Read entry at data table address
308
Latch contents
309
311
No match result
No
Valid?
310
Yes
312
MAC & VLAN match?
No
L2 link address
data address
314
Yes
313
Feed associated data to switching engine

FIG.4

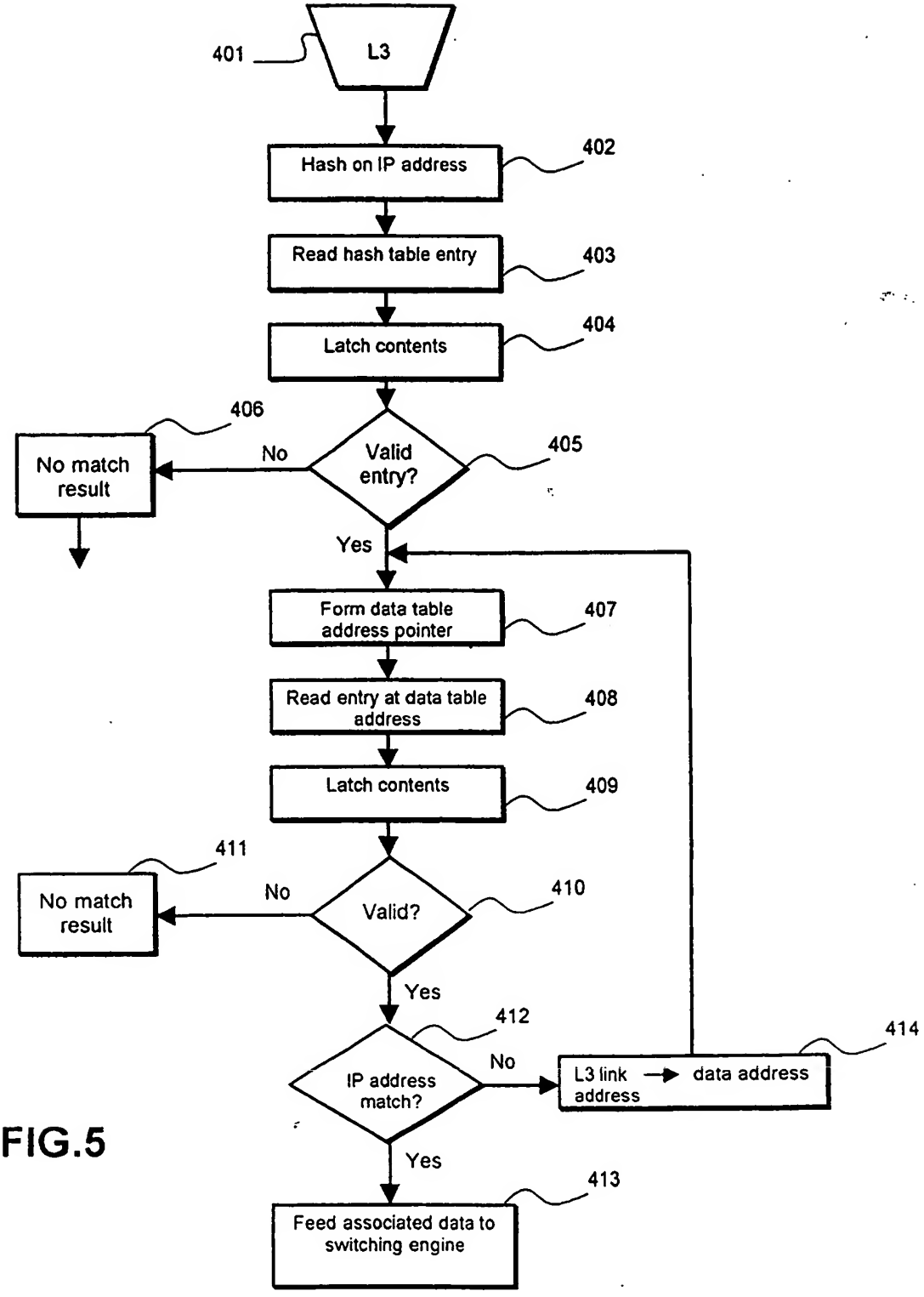

401
L3
Hash on IP address
402
Read hash table entry
403
Latch contents
404
406
No match result
No
Valid entry?
405
Yes
Form data table address pointer
407
Read entry at data table address
408
Latch contents
409
411
No match result
No
Valid?
410
Yes
412
IP address match?
No
L3 link address
data address
414
Yes
413
Feed associated data to switching engine

FIG.5

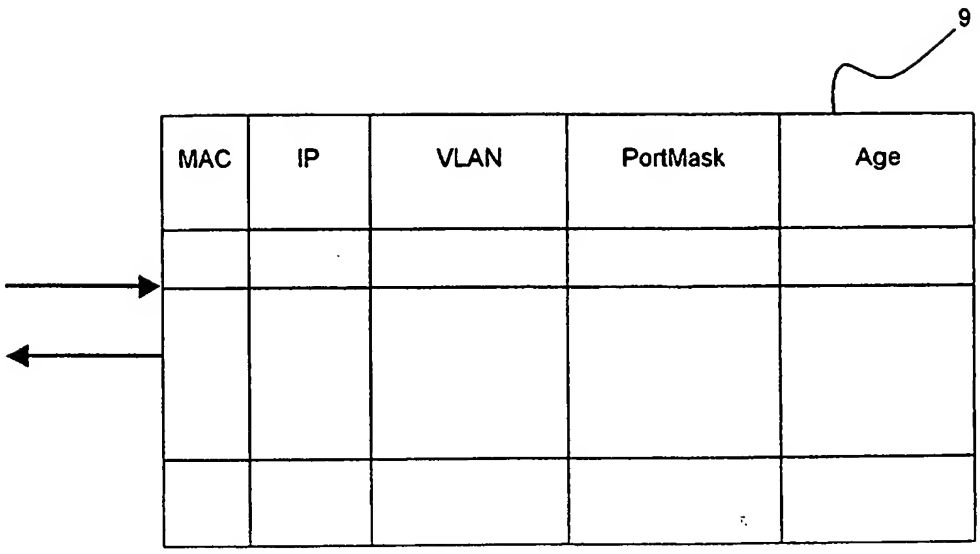

9
MAC
IP
VLAN
PortMask
Age

FIG.6

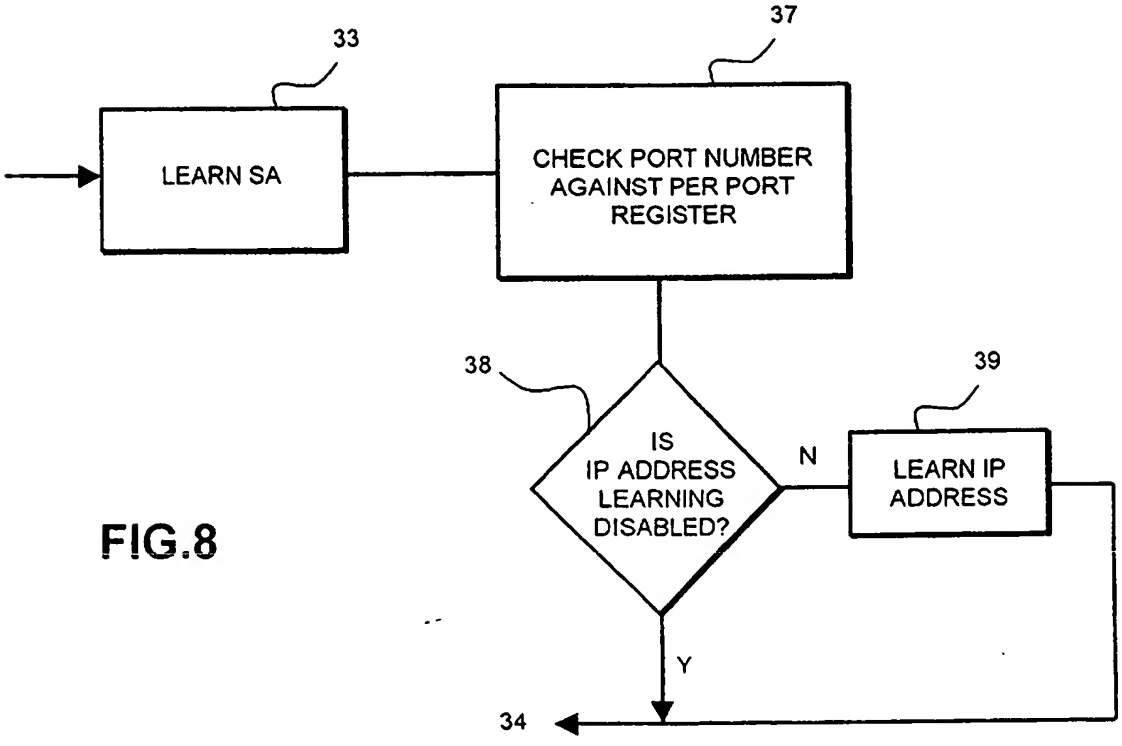

33
37
LEARN SA
CHECK PORT NUMBER AGAINST PER PORT REGISTER
38
39
IS IP ADDRESS LEARNING DISABLED?
N
LEARN IP ADDRESS
Y
34

FIG.8

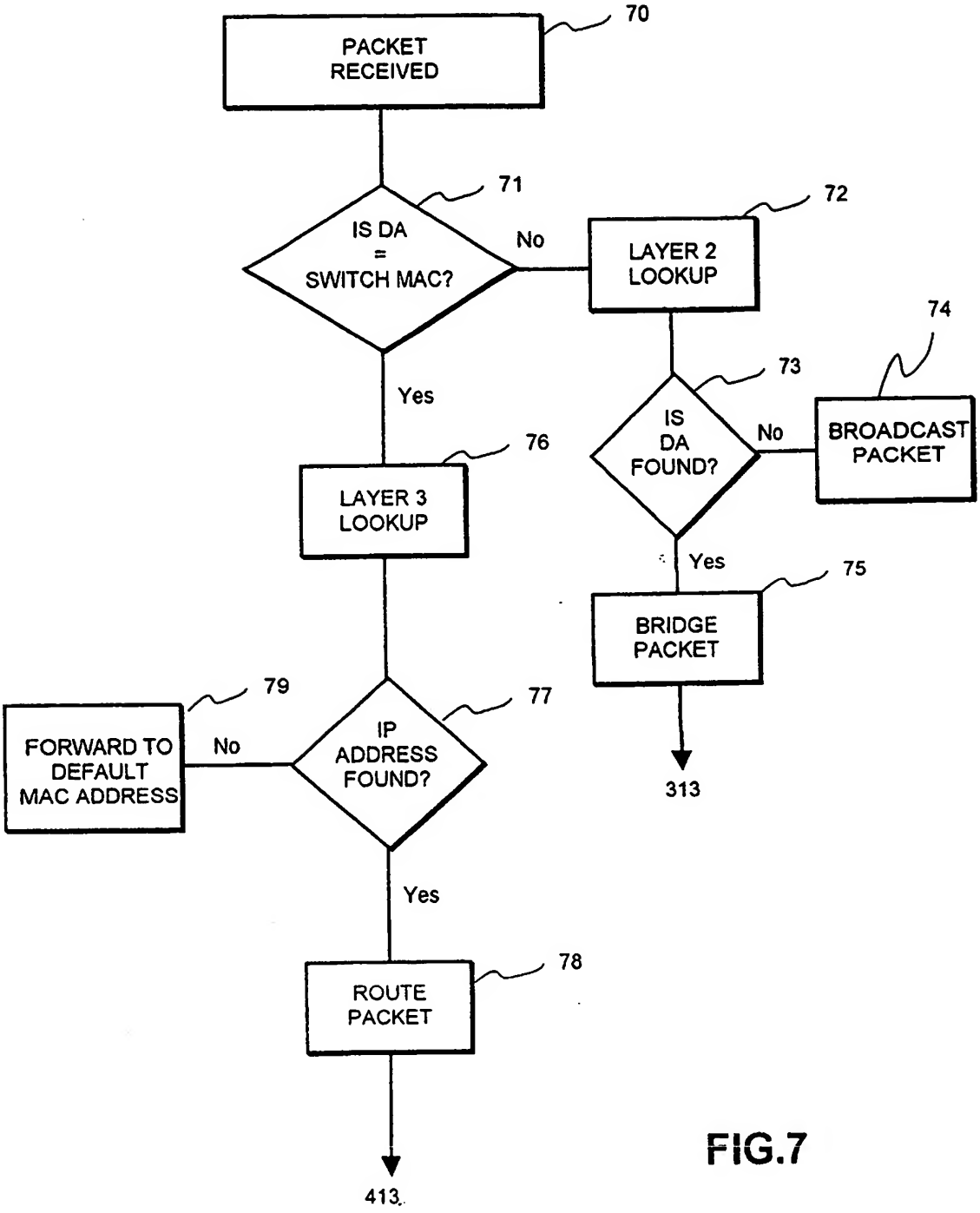

PACKET RECEIVED
70
IS DA = SWITCH MAC?
71
No
LAYER 2 LOOKUP
72
Yes
IS DA FOUND?
73
No
BROADCAST PACKET
74
Yes
BRIDGE PACKET
75
313
LAYER 3 LOOKUP
76
IP ADDRESS FOUND?
77
No
FORWARD TO DEFAULT MAC ADDRESS
79
Yes
ROUTE PACKET
78
413

FIG.7

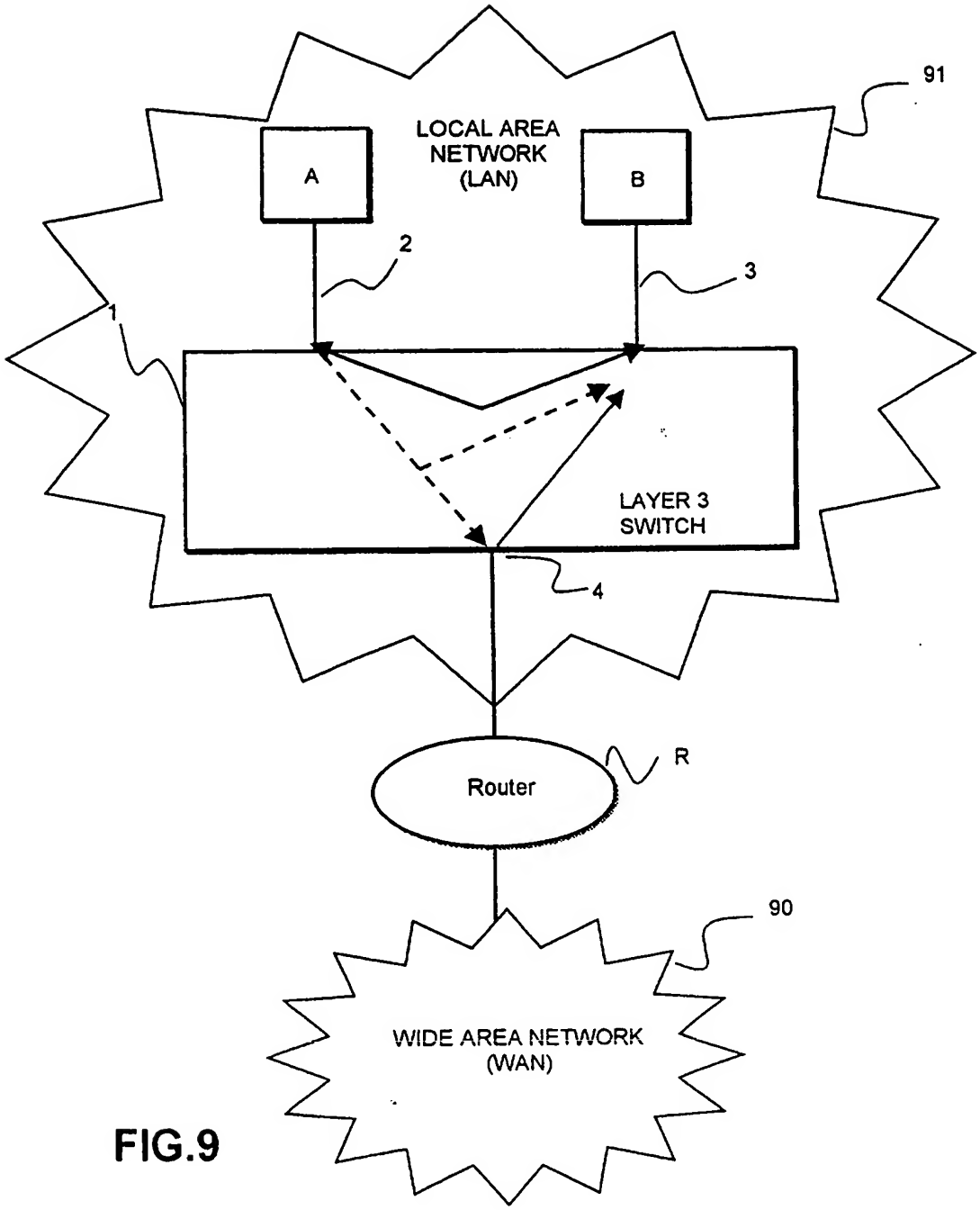

91
LOCAL AREA
NETWORK
(LAN)
A
B
2
3
1
LAYER 3
SWITCH
4
R
Router
90
WIDE AREA NETWORK
(WAN)

FIG.9

2358760

NETWORK SWITCH WITH SELF-LEARNING ROUTING FACILITY

Field of the Invention

The present invention relates generally to packet-based communication systems wherein data packets including address data and message or control data or both are propagated about a network in accordance with address data in the packets. The invention particularly relates to a network switch which includes a forwarding database and a multiplicity of ports of which one is connected to a router. The invention is mainly intended to facilitate the insertion into a network of a switch which can respond to protocol addresses and be able to relieve the router of traffic which requires simple routing decisions.

Background to the Invention

As is well known, a data packet is typically formed in a relatively high level of a communications protocol and before it is transmitted from an originating device it has attached to it a header which includes address data. The address data normally includes a protocol or network address, defining a 'subnet' on which the destination station is located and usually also an identification of that destination station. The network layer or protocol address is employed by a 'router', which term is intended to include devices which have a routing facility, to direct the packet to the appropriate subnet.

The address data within a packet needs to include at some stage a media access control address, otherwise known as a 'layer 2' or 'data link' address. The media access control (MAC) address is employed by switches and other devices to determine, when forwarding a packet, the specific device to which the packet should be sent.

When a packet does not contain a media access control address, as when for example a first end station initially attempts to communicate with a destination end station, it is necessary to perform an address resolution protocol, wherein a packet is broadcast indiscriminately. An end station receiving an address resolution packet (ARP packet)

containing its protocol address can reply with its media access control address. That enables a switch to establish in a forwarding database an entry which relates the particular protocol address with a media access control address and typically a port number of a port to which packets addressed to that destination end station will be sent.

Some operations in response to address data in packets are comparatively simple and speedy. For example, 'bridging' is the conventional term employed for responding to the MAC address and directing a packet to the device identified in that address. However, other forms of response, particularly 'routing' are more complex and require greater time. In particular, it is necessary to employ a router to perform such tasks as determining a best route for a packet to take, the prevention of indefinite looping of a packet, and a variety of other functions well known to those skilled in the art. Commonly, the performance of those ancillary functions is very much slower than the performance of a bridging function.

However, it is known to provide a switch which can operate both in 'layer 2' and 'layer 3', and which more particularly can in a default mode perform layer 2 look-ups but which can act also as a router, performing layer 3 look-ups. One example of such a switch, which operates with a single look-up table for both bridging and routing decisions, is disclosed in published GB patent application No. 2337674. Another example is a 12 port 100/1000 Mb/s Ethernet switch type 3C17700 made by 3Com Corporation. Such switches require configuration of their routing tables in order to operate in 'layer 3' but differ from fully functional routers in that the routing can be performed at high speed, e.g. 'wire speed'. It is presumed in the following that the 'network switch' employed is a switch of this character, being capable of 'layer 2' switching and, with appropriate configuration, 'layer 3' switching. Such a switch does not perform all the functions which a software controlled router can normally perform. If such a network switch is employed for example in a local area network and a router is also provided to route packets between that local area network and (for example) a wide-area network it would be beneficial to offload from the router the basic routing function (layer 3 switching) within the local area network and in particular, irrespective of what kind of router is employed, to employ the layer 3 switching capability of the switch to route local traffic between subnets.

As will be apparent to those skilled in the art and as more particularly discussed hereinafter, a switch which is capable of 'layer 3' routing will include in its forwarding database entries which relate a protocol (IP) address, a MAC (media access control) address identifying the next hop of a package intended for the protocol address and an identification (such as a number or port mask) of the port to which a packet routed according to that entry must be sent by the switch. The action of establishing in a database an entry of that character (which may include an identification of a virtual local area network) is usually termed 'learning' the protocol address for a particular port.

It is known in itself to control a switch, for example remotely by way of a 'management' port, so that protocol (layer 3) addresses cannot be 'learned' for a particular port.

Summary of the Invention

The main object of the invention is to facilitate the offloading of routing decisions from a router to an associated switch which is capable of switching decisions on both media access control address and protocol addresses. The main feature of the invention is to provide the connection between the router and the switch only by way of a port in respect of which the switch can learn media access control addresses but is unable (for example by being specifically disabled) to learn protocol (IP) addresses.

Further advantages and features of the invention will become apparent from the following detailed description with reference to the accompanying drawings.

Brief Description of the Drawings

Figure 1 is a general schematic representation of a known form of switch which may be employed in the invention.

Figure 2 is a simplified illustration of an addressed data packet.

Figure 3 is a flow diagram illustrating a learning process for a network switch.

Figure 4 is a flow diagram of principally a 'layer 2' look-up process in a network switch.

Figure 5 is a flow diagram of a 'layer 3' look-up process in a network switch.

Figure 6 is a simplified schematic illustration of a data table.

Figure 7 is a flow diagram of the operation of a switch capable of layer 2 and layer 3 switching decisions.

Figure 8 is a partial flow diagram of an address learning process.

Figure 9 is a schematic illustration of a connection of a switch and a router in accordance with the invention.

Detailed Description

Although the specific instruction of a switch is not necessarily an important feature of the invention, provided that the switch has both the storage ability and the processing ability that the invention requires, Figure 1 is intended to show schematically the basic components of a switch that is suitable for use in the present invention. Typically, switches have twelve or twenty-four ports or even more. For the sake of simplicity, the switch 1 shown in Figure 1 has only four ports, identified as ports 2, 3, 4 and 5. As will be seen later, it will be assumed that ports 2 and 3 are connected to other network devices, port 4 is connected to a 'router' and port 5 is a management port by means of which the switch can be configured by remote control in a manner well known to those skilled in the art.

If, as is preferred, the switch 1 is primarily a hardware switch, the various components within the switch 1, apart from most of the memory, be provided on a single ASIC

(application specific integrated circuit). However, for ease of explanation, the various components of the switch are separately shown in Figure 1. In this example therefore, each of the ports 2, 3, 4 and 5 has a respective 'port ASIC', 2a, 3a, 4a and 5a respectively. These components include the media access control devices (MACs) which perform (known) operations on packets entering and leaving the switch while the packets are in a format independent of the particular transmission medium to which a respective port is connected. The port ASICs also include a 'physical layer device' which not only converts packets from a media independent format to a format appropriate for the particular transmission medium but also includes various other functions such as for example auto-negotiation, particularly in the case of 'Ethernet' networks described in IEEE Standard 802.3.

The switch 1 includes a bus system 3 by means of which packet data and control and status data are conveyed between the various components of the switch. The switch includes a look-up engine 7, the operation of which will be described later, a memory 8 which may be employed for the temporary storage of packets in 'queues' before they are sent to their destination ports, a forwarding database 9, which will be described with reference to Figure 6, and a switching engine 10. The switching engine will retrieve packets temporarily stored in memory 8 and direct them to respective ports in accordance with, for example, a port mask obtained from a relevant entry in the forwarding database 9. The switch also includes a register 11 the function of which will be explained later.

Figure 2 illustrates in simplified schematic form a typical packet employed for the conveyance of data in a packet-based data communication system in which a switch such as switch 1 may form part. The packet comprises a start-of-frame delimiter (SFD), media access control address information, comprising a destination address (DA) and a source address (SA), protocol data, message data and cyclic redundancy check data (CRC). The media access control addresses define, if they are present, the source and destination devices in one 'hop' of a packet. The protocol data includes network address data defining, for example, the network to which the ultimate destination of the packet belongs

and usually also an identification of a device within that network. The message data need not be present, as in the case of a control packet.

Figure 3 illustrates mostly the learning process for MAC addresses typical of a network switch. A packet is received, stage 31, and a look-up, performed by means of look-up engine 7 in forwarding database 9, determines whether the source address (SA) is already the subject of an entry in the database. If it is not, then the address is 'learned' (stage 33), that is to say made the subject of a table entry including an identification of the port on which the packet was received and a VLAN number. If the switch is to be used for routing (layer 3 switching) as well as bridging (layer 2 switching), an entry will typically include the protocol (IP) address of the packet.

In ordinary, layer2/layer switches, IP addresses may be learned at this stage. The switch 1 includes on its ASIC a per port register 11 which identifies those ports for which IP addresses may not be learned. This will be further explained with reference to Figure 8.

In order to determine where the packet should be sent, a further look-up is made (stage 34) to find a match for the destination address (DA) in the database. If the address is found, then the packet may be forwarded (stage 35) from the port associated with that MAC address in the forwarding database. For this purpose the entry is read out from the forwarding database and fed to the switching engine 10.

If it should happen that the destination MAC address is not in the forwarding database, it is normally necessary to 'flood' or 'broadcast' the packet (stage 36). By this is meant that a copy of the packet is supplied to all (or all of a selected plurality) of the ports in order to obtain an ARP (address resolution protocol) response from a device having the network address identified in the packet. That device will respond with its MAC address and enable this address to be learned in respect of the relevant port in the forwarding database.

Figure 3 (particularly stage 34) is intended to include the case when the MAC destination address (DA) of the packet matches the MAC address of the switch; if the packet is of

appropriate IP type, it can be routed (stage 35). If the destination IP address is not in the database the packet would be sent by a default route, but not broadcast as in the case of an (unsuccessful) layer 2 look-up.

Figures 4 and 5 will be discussed in relation to Figure 6, which illustrates a typical 'combined' data table which by way of example may perform the functions of a 'routing table', a 'bridging table' and an 'ARP cache'. This shows a forwarding database which contain a multiplicity of entries which may contain a MAC address, a subnet or VLAN address, a network (IP) address, a port mask and an age field. This database is accordingly organised as described in the aforementioned GB-A-2337674.

Figures 4 and 5 illustrate the manner of performing look-ups in a forwarding database. These Figures correspond to Figures in the aforementioned GB-A-2337674. Both Figures assume that the process of look-up is facilitated by means of the hashing of the address which is the subject of the look-up (whether this be a combination of the destination address and VLAN address or the IP address).

Referring first to Figure 4, stage 301 illustrates a decision stage determining whether the MAC address is within a local range of MAC addresses. If the MAC address (DA) is within that local range but the packet type is not IP, then the packet must be bridged. The decision process associated with decision stage 301 will be described with reference to Figure 7.

If a layer 2 look-up is to be performed, the switch will perform a hash operation on a combination of the packets destination address (DA) and VLAN number, stage 302, the hash table entry is read (stage 303) and the contents latch (stage 304). The next stage (305) is an examination whether the entry is valid. This need mean no more than a determination whether the entry is still current or has been aged. If the entry is not valid, then the search fails (stage 306). If the entry is valid, then a data table address pointer is formed from the latched contents of the hash table entry, stage 307, the entry is read, stage 308, the contents of the entry are latched, stage 309, and a determination whether the entry is valid

is made (stage 310). If the entry is not valid, then no result has been obtained, stage 311. If the entry is valid then it is determined, stage 312, whether the MAC address and VLAN number in the entry match those of the destination address which is being looked up. If they do not, then it may be necessary to search another entry, linked to the first by means of a pointer. This expedient is necessary because the hashing of addresses may mean that a plurality of addresses may hash to the same entry. However, both the hashing of addresses, and the use of link pointers are merely preferred features of the switch described in the aforementioned prior application and are not essential to the present invention. The important matter is whether a look-up has found the destination address in the forwarding database. If it has been, the response to stage 312 being 'yes' then the data associated with the entry (such as the port mask) are fed to the switching engine, stage 313.

Figure 5 illustrates a similar look-up which may be performed in respect of layer 3 addresses. In this case, the entry stage is stage 401. The IP address is hashed (402), the hash table entry is read (403) and the contents thereof latched (404). On examination (stage 405) of the validity of the entry, there is no match result if the table entry is invalid (406). If the table entry is valid then a data table address is pointer is formed (stage 407), the entry at the data table address is read (stage 408), the contents latched (409) and a test of validity made (stage 410). If the table entry is not valid then there is no match result (stage 411). If the table entry is valid then there is a test to determine whether the IP addresses match, stage 412. If there is a list of addresses linked by pointers, because of the use of hash tables, then a link pointer 414 points to another address in the table and the loop from stage 414 via 407 to stage 412 is reiterated.

Again, however, the important matter is whether a match of the IP address in the packet has been found with an address in the data table. If so, then the relevant data in that entry including the port mask is fed to the switching engine, shown in stage 413.

Figure 7 is a summary of the decision process in stage 301 as well as a summary of the layer 2 and layer 3 look-ups shown in Figures 4 and 5. By way of introduction to Figure 7,

it should be remarked that the switch 1 is normally, for example by way of the management port 5, configured with its own MAC address.

Referring now specifically to Figure 7, if a valid IP packet, that is to say a packet containing a valid 'network' or layer 3 address, stage 70, a determination (stage 71) is made to

If the incoming valid IP packet does not contain the local MAC address, then the response to stage 71 (which corresponds to stage 301 in Figure 4) is negative and the switch will perform a layer 2 look-up, summarised by stage 72 and more particularly illustrated in Figure 4. If the destination MAC address is found then the packet may be 'bridged', that is to say switched to the relevant port on the basis of the MAC address and port number. If the MAC address is not found in the database then the packet will be broadcast, stage 74.

If however, the incoming packet does have a destination MAC address (DA) corresponding to the address of the switch, a layer 3 look-up will be performed (stage 76 and Figure 5). If the layer 3 (network or IP) address is found by for example the process shown in Figure 5, the search will retrieve the next hop MAC address and therefore the relevant port number and the packet can be 'routed'. It will, in well known manner not requiring description, have its MAC source address changed to the MAC address of the switch 1 and the 'TTL' will be decremented.

Figure 7 shows a further situation, where the IP address search, stage 76, yields a negative result. In this case, the packet needs to be forwarded by a default route. This is defined by a 'default' MAC address.

The foregoing description is intended to provide the reader with a background for understanding the learning and look-up processes in a layer 2/layer 3 switch of the kind which may form a combination with a router according to the invention.

Disabling of Learning

It is known, for example, in a 'local office interconnect' scheme to modify the operation of a switch such as switch 1 by preventing the learning of IP addresses in respect of a selected port.

As indicated below the layer 3 switch router listens on the network for 'router alive' messages. When it detects such a message it reads the source port number that the source MAC address has been learned against. It then writes that port number to a per port register which disable learning for that port. It does not disable MAC address learning so MAC address can continue to be added to the layer 2 database.

'Router alive' messages are transmitted at various intervals and can be detected by snooping on, for example OSPF, PIM and RIP packets.

It is also feasible to disable learning for a particular port by local (e.g. manual) programming of the switch.

Whether the learning is disabled automatically or not, the ability to learn IP addresses against a port is controlled by the 'per port' register 11.

The learning of IP addresses is illustrated in Figure 8, which is intended to be read in conjunction with Figure 3. Thus, although the switch 1 learns MAC addresses routinely (stage 33), the learning of the IP address of a packet depends on a check (stage 37) of the port number of the incoming packet against the per port register 11. If IP address learning is disabled for that port (stage 38) the learning process reverts to stage 34 of Figure 3. This does not preclude layer 3 switching (routing) if the MAC DA of the packet matches the MAC address of the switch. If IP address learning for that port is not disabled, the IP address is learnt against the respective port (stage 39).

Offloading of a Router

Reference will now be made to Figure 9 which shows a layer 2/layer 3 switch 1 as described in the foregoing connected by way of port 2 to a first network device A, connected by way of port 3 to second network device B and by way of port 4 only to a router R. As described in the foregoing, the switch 1 is configured so that it is unable to learn IP addresses in respect of port 4.

In this typical example, the router R is a 'standard' router which performs, principally under software control, wide area routing functions for a wide-area network 90. The switch 1 is principally intended for operation on a local-area network (LAN) 91 to route traffic between sub-nets of the LAN. Large routers such as router R must perform considerably more packet analysis than a switch 1 and not all their functions can (unlike switch 1) be implemented in hardware (i.e. in an ASIC). The router R may route traffic between the LAN 91 and WAN 90 but should preferably not route traffic within the LAN (e.g. between stations A and B).

In a first phase, it will be assumed that the switch 1 receives an ARP packet from station A. The switch will recognise such a packet as not an ordinary IP packet, since ARP packets have a different type field within the packet. The switch 1 will forward such an ARP packet to every possible destination, including station B by way of port 3 and the router by way of port 4. It will also be assumed that an ARP response is generated by station B. This response will have the MAC address of station B and this response will be sent by way of switch 1 to the router R. The router will forward the ARP response packet back to station A together with the MAC address of the router.

At this point no 'layer 3' IP addresses will have be learned by the switch 1 though layer 2 addresses will have been learned through standard bridging rules. In particular, the router's source MAC address will have been learned in forwarding database 9 when the router replies by sending the ARP response packet back to station A.

In a second phase, an IP packet is sent by station A. The switch will perform a look-up in respect of this IP packet. On the assumption that the IP source address of station A is not

in the database, that IP source address will be entered in the forwarding database 9 in respect of port 1.

The switch 1 will perform a further look-up for the destination IP address in the packet. It will find that the address is not in the forwarding database 9 because the IP address of station B has not yet been learned. Accordingly the switch will forward the packet to the router, this being the 'default' address illustrated in stage 79. A default MAC address is necessary according to standard routing rules.

In a third phase, the router will route the IP packet generated by A on to station B. The destination protocol address of B will not be in the table but there will be no learning of the IP source address (station A) for port 4 because it has been disabled in respect of the learning process.

Finally, station B will send a packet to station A by way of reply. The switch will find the destination of address of station A in the database and according the packet will be routed by the switch to station A. The switch will learn the source's protocol and MAC addresses for station B in the database and from then on station A can communicate with station B without the aid of the router.

Claims

1. A sub-system for forming part of a packet-data communication system for the transmission of data packets which contain media access control addresses and protocol addresses, the sub-system comprising:

a multiple-port network switch for forwarding data packets in response to an examination of the packets on the basis of the packets protocol addresses, and for performing said examination when a destination address in the packet corresponds to the media access control address of the switch, the switch being arranged for forwarding a packet on the basis of a media access control address therein when the destination address of the packet does not correspond to the media access control address of the switch, said switch including a forwarding database which comprises entries which relate a protocol address with a media access control address and an identification of a port of the switch;

wherein the switch is configured so that protocol addresses may not be learned in the database in respect of said particular port; and

a network router which is connected to the switch only by way of said particular port.

2. A method of relieving traffic flow in a network router which forms part of a packet-based data communication system in which data is conveyed by means of addressed data packets including media access control addresses and protocol addresses, said method comprising:

connecting to the router a network switch which is capable of switching data packets from port to port both in response to media access control addresses and protocol addresses and includes a database for containing entries relating protocol addresses to particular ports;

disabling the database in respect of a specific port so that protocol addresses cannot be learned by the database in respect of that specific port;

said router being connected to the switch only by way of the specific port.

3. A method according to claim 2 wherein said switch is disposed to route data packets within a local-area network and said router is disposed to route data packets within a wide-area network.

The Patent Office
PATENTS • DESIGNS • COPYRIGHT • TRADE MARKS
INVESTOR IN PEOPLE

Application No: GB 0001654.3
Claims searched: 1-3

Examiner: Richard Howe
Date of search: 28 June 2000

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.R): H4K (KTK) ; H4P (PPS)

Int Cl (Ed.7): H04L (12/56) ; H04Q (11/04)

Other: Online : wpi ; epodoc ; japio

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
A	GB 2 337 674 A (3Com) - see whole document	
A	GB 2 337 659 A (3Com) - see whole document	
A	EP 0 835 009 A2 (Kabushiki Kaisha Toshiba) - see whole document	

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.